Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017)
IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9

Predicting Movie Box Office Success using Multiple Regression and SVM

Subramaniyaswamy V., Viginesh Vaibhav M., Vishnu Prasad R., Logesh R.
School of Computing,
SASTRA University,
Thanjavur-613401, India
vsubramaniyaswamy@gmail.com

***Abstract*— Hollywood is a multi-billion dollar industry which releases more than a hundred films a year, with large variations in the budgets and box office grosses of the movies. Identifying which factors are important to a movie's profitability and subsequently predicting the success of a movie given its relevant parameters could save movie studios hundreds of millions of dollars a year. This paper analyses the efficiency of using multiple linear regression and Support Vector Machine Classification to predict the box-office success of movies, while analysing the influence of variables like trailer views, Wikipedia page views, critic ratings and time of release.**

Keywords— Data Mining; Data Analytics; Prediction; Box Office Gross; Regression; SVM Classification; Machine Learning

I. Introduction

In 2015, the global box office gross reached an all-time high of $38 billion, with 5 films grossing over 1 billion, the most in the history of Hollywood [1]. Most of the highest grossing films came from just 6 studios: 20th Century Fox, Marvel Studios, Walt Disney Pictures, Columbia Pictures, Paramount and Warner Bros. However, the total gross of a movie is not always indicative of its profitability. Because of skyrocketing costs of marketing campaigns attached to blockbusters with a budgets crossing $200 million, movies with a larger initial budgets often need to gross a higher percentage of their budget to break even. Hence, there is a need for an adjustable ROI (Return On Investment) that differentiates between smaller films and blockbusters.

The ROI is influenced by a variety of factors, many of which are available publicly on the internet. Because of this, there is great potential for using the extensive information available to predict the success of movies. First, we collect the relevant data from the internet and clean it to remove junk values and discard movies which do not have available information about them. The data is in both numerical and nominal form. They are stored is Comma Separated Variable (CSV) format for easy analysis. Then, we look at what factors most impact a films adjusted ROI, and to what extent. Finally, we predict a movies success or failure based on the contents of its relevant parameters, using multiple linear regression, and Support Vector Machine Classification in Weka.

II. Litrature review

Past attempts to predict movie success with already available information has generally focused only on a single variable. Marton Mestyan et al. [2] studied at the impact of Wikipedia on movie success and looked at number of views, number of users, number of edits and collaborative rigor, to conclude that Wikipedia was a better indicator of success than Twitter a month before a movie's release (Coefficient of determination R>0.95), whereas Twitter was a better indicator a few days before a movie released(R>0.98). Asur and Huberman [3] used Twitter activity to train and test their model, but limited it to activity only just before the movie released. Similar studies have also had only limited success [10].

Dan Cocuzzo and Stephen Wu of Stanford University [4] tried to predict Box office gross using IMDb ratings, making of use of Naïve Bayes and Support Vector Machine classification methods. They found that both classifiers excelled at finding highly unprofitable movies, but weren't very useful for predicting hits.

Another study by Nithin VR et al. also looked at IMDb ratings while using Linear Regression, Logistic Regression and SVM classification to predict success. They found that the three methods had a success rate of 50.7%, 42.2% and 39% respectively, thus indicating that linear regression was the best predictor. [5]. Michael T. Lash and Kang Zhao found that the "average profit of previous actor-director collaboration" was the most important factor for the success of a movie [6].

Most previous efforts take into consideration a single factor only and try to predict what impact it has on final Box office gross. Not many studies have looked at the impact of and trailer views on movie success, or critic reviews using critic aggregation sites like Rotten Tomatoes.

III. Data collection

The dataset was populated with information scraped from BoxOfficeMojo and Wikipedia for movies released in 2016. Trailer views were taken from YouTube. A simple script enabled content in tables from Wikipedia page to be pasted to a Google Sheets file, from where it could be exported to Excel. The features collected include the following:

978-1-5386-1959-9/17/$31.00 ©2017 IEEE

- Opening Date
- Movie Name
- Budget
- Domestic Gross
- International Gross
- Total Gross
- Trailer Views
- Studio
- Cast and Crew
- Genre
- Medium (Live action or Film)
- Trailer Views
- Wikipedia Views
- Rotten Tomatoes Score

A. Data Pruning

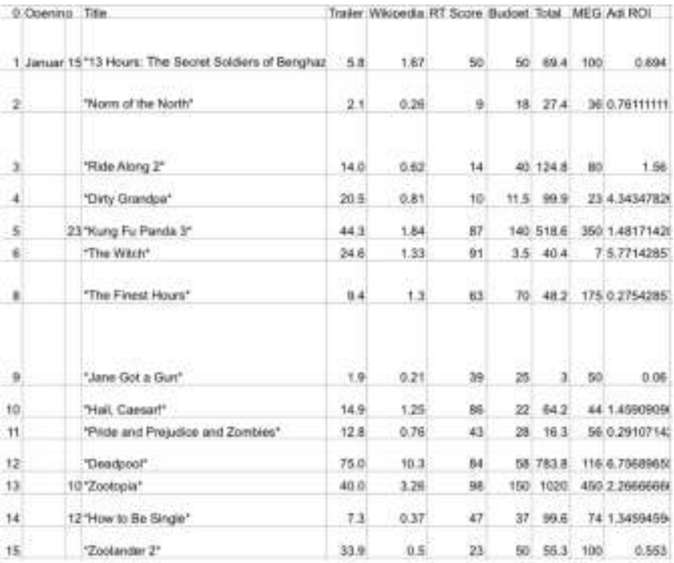
From the resultant dataset, movies which had incomplete information or junk values had to be removed. The vast majority of the pruned movies were low budget films released in a limited run and for whom information wasn't publicly available. Next, duplicate entries had to be removed which was a consequence of an overlap between movies from Wikipedia and BoxOfficeMojo. From this list, movies with very low trailer views and non-existent Wikipedia pages had to be removed as our analysis considers both of these factors when making a prediction. The majority of the affected films were once more low budget indie releases. At the end of the pruning process, 138 films were left that had complete data. These films were mostly released by major studios.

B. Data Classification

We classified the movies in the dataset as part of 3 categories:

- Low Budget Films (Budget < $50 million)
- Medium Budget Films (Budget between $50 million and $150 million)
- Big Budget Films (Budget >$150 million)

Each of these categories has a different multiplier to be considered a success: x2 for Low Budget Films, x2.5 for Medium Budget Films and x3 for Big Budget Films. This is to account for the increased marketing costs as the budget increases, meaning the studio needs a larger ROI to break even. For example, a movie with a budget of $20 million that grosses over twice its budget at $40 million would be considered a success, while a movie with a budget of $160 million would need more 3 times its budget ($480 million) in order to be considered a success. We use this multiplier to get a Minimum Expected Gross (MEG) which is the threshold a film needs to cross to be considered a success. Finally, we define a variable called Adjusted Return On Investment (Adjusted ROI) which is a measure of the success of the movie.



0	Opening		Title	Trailer	Wikipedia	RT Score	Budget	Total	MEG	Adj ROI
1	Januar	15	"13 Hours: The Secret Soldiers of Benghaz	5.8	1.67	50	50	69.4	100	0.694
2			"Norm of the North"	2.1	0.26	9	18	27.4	36	0.76111111
3			"Ride Along 2"	14.0	0.62	14	40	124.8	80	1.56
4			"Dirty Grandpa"	20.5	0.81	10	11.5	99.9	23	4.3434782
5		23	"Kung Fu Panda 3"	44.3	1.84	87	140	518.6	350	1.4817142
6			"The Witch"	24.6	1.33	91	3.5	40.4	7	5.7714285
8			"The Finest Hours"	8.4	1.3	63	70	48.2	175	0.2754285
9			"Jane Got a Gun"	1.9	0.21	39	25	3	50	0.06
10			"Hail, Caesar!"	14.9	1.25	86	22	64.2	44	1.4590909
11			"Pride and Prejudice and Zombies"	12.8	0.76	43	28	16.3	56	0.2910714
12			"Deadpool"	75.0	10.3	84	58	783.8	116	6.7068965
13		10	"Zootopia"	40.0	3.26	98	150	1020	450	2.2666666
14		12	"How to Be Single"	7.3	0.37	47	37	99.6	74	1.3459459
15			"Zoolander 2"	33.9	0.5	23	50	55.3	100	0.553

Fig. 1: Sample of the dataset with trailer views (in millions), Wikipedia views, RT Score Budget, Total Gross, Minimum Expected Gross (MEG) and Adjusted ROI

$$\text{Adjusted ROI} = \text{MEG}/\text{Budget} \qquad (1)$$

C. Data Distribution

We look at how data changed over the timeframe we collected for, to see if there are yearly patters for critic scores, Wikipedia views, trailer views and total gross. The distribution for Rotten Tomato (critic review aggregation site) scores is given in Fig. 2. It shows an increase towards the end of the year, where movies traditionally aiming for The Academy Awards are released [7]. Trailer views are shown in Fig. 3.

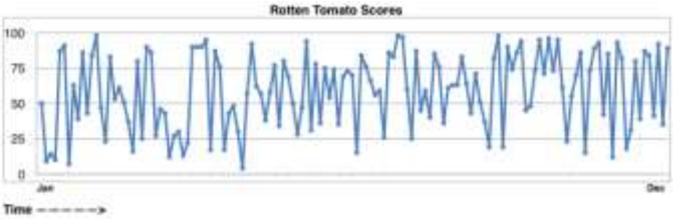


Fig. 2: Critic ratings variation throughout the year (0-100%)

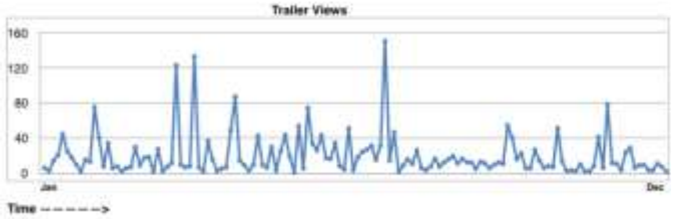


Fig. 3: Trailer view variations throughout the year in millions

Fig. 4 shows Wikipedia page views throughout the year. Trailer views and Wikipedia page views are both positively correlated, and both serve as a proxy for the public's interest in the movie weeks and months before release. Fig. 5 visualizes total box office grosses throughout the year. There is a significant drop off after the summer months. This is represented in Fig. 6, which depicts average monthly box office gross throughout the year. April-July are the months where the biggest grossing movies are released, with June being the biggest month in terms of revenue with an average of $269.16 million grossed per movie. Fig. 6 correlates with the budget, as movies with a big budget gross more, even if they don't recoup the studio's investment. Looking at the Average Adjusted ROI is a much more revealing measure of the success of a movie.

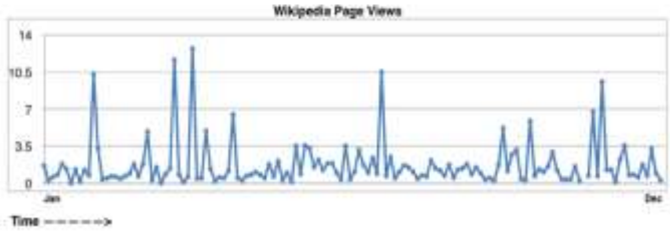


Fig. 4: Wikipedia page views throughout the year in millions

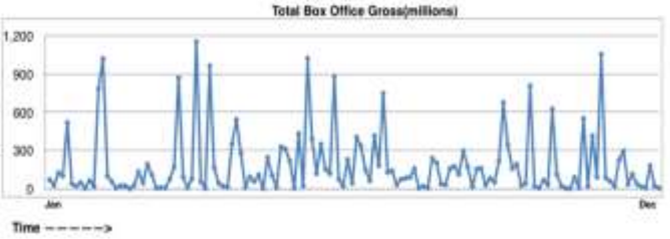


Fig. 5: Total Box Office Gross throughout the year in millions

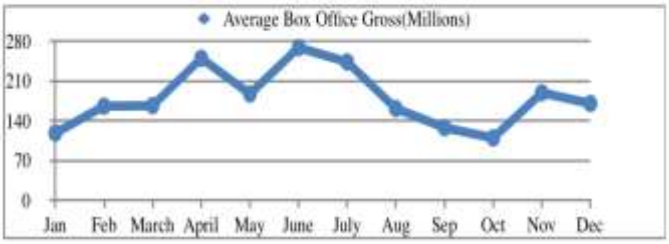


Fig. 6: Average Box Office Gross for movies by month

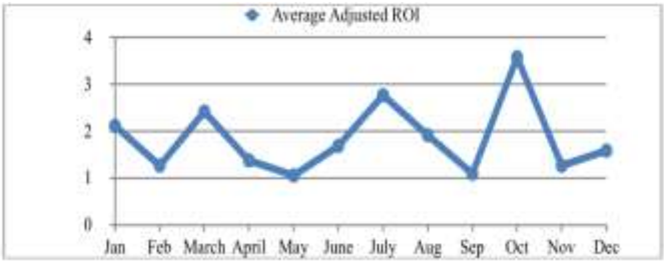


Fig. 7: Average Adjusted ROI for every month

Fig. 7 plots the Adjusted ROI over the months of the year, and finds that October gave the best results for movie studios, since well-performing small budget films released at that time grossed much higher than expected. For example, Moonlight (2016) was released in October and had a budget of $1.5 million. Its box office gross of $50 million made it the largest Adjusted ROI (33.3) amongst movies re-leased in 2016.

IV. Methods

A. Data Pruning

Linear regression analysis usually involves a formula of the form

$$Y=a+bX \qquad (2)$$

A multiple regression has multiple 'X' or independent variables in one formula. Multiple linear regression considers more than one independent variable, and can be described as a particular case of general linear models, when the number of dependent variables(Y) is one. The basic model for multiple linear regression is

$$Yi = b0 + b1Xi1 + b2Xi2 +.... + bpXip + ei \qquad (3)$$

We applied multiple regression to predict the success of the movie, with the independent variables being Trailer Views, Budget, Critic Ratings and Wikipedia Page Views. Where we consider n observations of one dependent variable and p independent variables b0, b1, etc. represent parameters to be estimated, ei is the i^{th} independent identically distributed normal error.

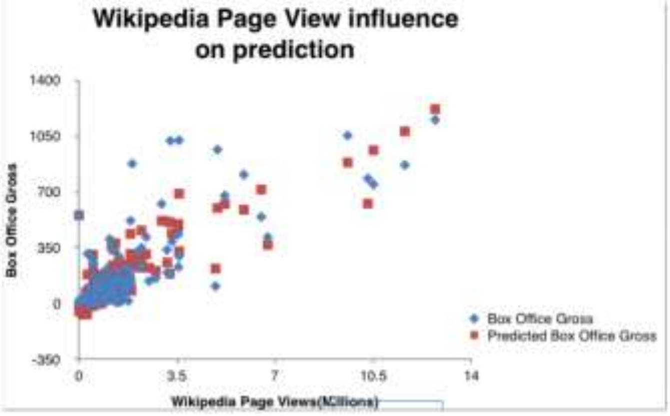


Fig. 8: Wikipedia Page Views (in millions) used to predict box office gross

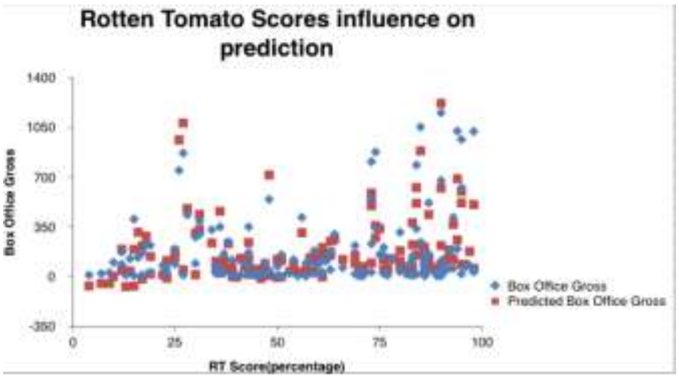
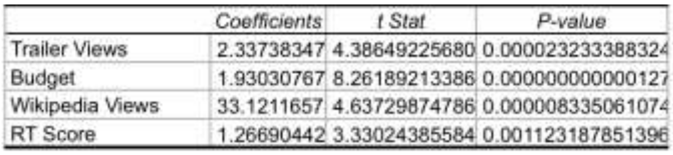


Fig. 9: Rotten Tomato Scores influence on total gross prediction

We then find the t-statistic, which is the ratio of the difference of the estimated value of a parameter from its expected value to its standard error, and p-value, which can be defined as the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event.



	Coefficients	*t Stat*	*P-value*
Trailer Views	2.33738347	4.38649225680	0.00002323338832
Budget	1.93030767	8.26189213386	0.0000000000012
Wikipedia Views	33.1211657	4.63729874786	0.00000833506107
RT Score	1.26690442	3.33024385584	0.00112318785139

Fig. 10: Comparison of coefficient value, t-stat and P-Value for independent variables

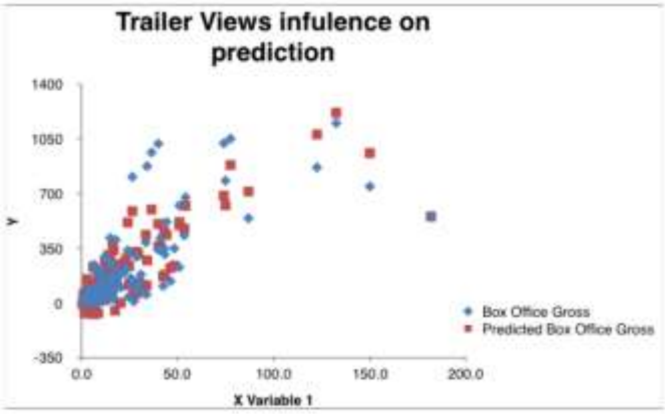


Fig. 11: Influence of trailer views on box office gross prediction

We see that budget has the highest t-Stat value of 8.26, and RT scores with lowest t-stat value (3.33) and highest p-value (0.0011) indicating that it has the least impact on accurate prediction, which is consistent with previous findings [11]. Further, when combing all independent variables and their coefficients to perform multiple regression, we get an R-value of 0.8893. Performing R square regression where independent variables are squared, we get the coefficient of determination R2 as 0.7909. Further results are shown in table 1.

Table 1: Experimental results of SVM

Correctly Classified Instances	78 (56.5217 %)
Incorrectly Classified Instances	60 (43.4783 %)
Kappa statistic	0.2384
Mean Absolute Error	0.3098
Root Mean Squared Error	0.396

As budget has the lowest p-value, it is the most accurate predictor of box office gross. But higher box office gross is not necessarily indicative of a better Return On Investment (ROI). To predict the best ROI factor, we use Support Vector Classification method (SVM) to train the relevant features to predict best ROI, as discussed in the next section.

B. *Support Vector Machine Classification*

Before beginning SVM classification, we need to group movies based on their Adjust ROI. We define 4 buckets into which all movies fall into:

- Adjusted ROI < 1 (Box office flop)
- Adjusted ROI between 1-1.5 (Barely breaks even)
- Adjusted ROI between 1.5-2.5 (Moderately successful)
- Adjusted ROI >2.5 (Hugely Successful)

SVMs, also known as Support Vector Networks, are supervised learning models used for regression analysis and classification. The input to the SVM is a training set, which can be a set of examples, where each example is associated with a particular category. SVM then creates a new model that assigns given examples to a different category (or the same one). It is described as a non-probabilistic binary linear classifier

In SVM, the models consist of points in space, where each point represents an example. Mapping is done in a way that tries to maximize the difference between examples from different classes, so that the examples of a particular class are very different from examples from another class. New examples are then as-signed to a class based on whichever side they are closer to. Making the class or category have high intra-class similarity but low inter-class similarity is a goal of SVM. We use 10-fold cross-validation for our dataset.

Kappa statistic is a measure of inter-rater agreement for qualitative items. It is a better measure than percentage agreement classifiers, since κ takes into account the possibility of agreements occurring by chance. Kappa statistic for the dataset is 0.2384, while the SVM correctly classified 56.52% of the movies, which is an improvement from previous attempts [5] [8] which only considered IMDb data and had an accuracy of 39%, thus with an improvement in accuracy of over 42% when accounting for trailer views, Wikipedia page views, RT score and budget over using only IMDb data.

V. Conclusions

Most past efforts at trying to predict movie box office success have focused only on one feature and its impact on movie success. We have taken into consideration different sets of publicly available relevant data and used multiple regression to come up with a predictor for movie success, with the resultant R-value greater than 0.88. This information could be used by movie studios to ad-just their marketing budget, which can be a significant drain on resources, so as to increase their ROI. Adjusted Return On Investment also shows that smaller budget films with good word-of-mouth often have a larger ROI than big budget block-busters.

SVM was used to train multiple variables and the result was an accurate classification rate of 56.52%, which is a higher accuracy than previous attempts at SVM using only a single variable.

ACKNOWLEDGMENT

The authors are grateful to Science and Engineering Research Board (SERB), Department of Science & Technology, New Delhi, for the financial support (No. YSS/2014/000718/ES).

REFERENCES

[1] Forbes (2016) "Experts Predict a Drop in Box Office Revenue In 2016 After a Record Year for Hollywood", https://www.forbes.com/sites/simonthompson/2016/01/05/experts-predict-a-drop-in-box-office-revenue-in-2016-after-a-record-year-for-hollywood/ #402059897195

[2] Mestyan M, Yasseri T, Kertesz J (2013) Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLoS ONE 8(8): e71226. doi:10.1371/journal.pone.0071226

[3] Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. pp. 492–499.

[4] Dan Cocuzzo, Stephen Wu, Hit or Flop: Box Office Prediction for Feature Films, Stanford University, 2013

[5] Nithin VR, Pranav M, Sarath Babu PB, Lijiya "A Predicting Movie Success Based on IMDB Data" International Journal of Data Mining Techniques and Applications ISSN: 2278-2419 Volume: 03, June 2014, Pages: 365-368R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Michael T. Lash and Kang Zhao Early Predictions of Movie Success: the Who, What, and When of Profitability, arXiv:1506.05382v2 [cs.AI] 29 Jan 2016

[7] New York Times, "There is more to winning an Oscar than meets the eye", (January 2015), https://www.nytimes.com/2015/01/29/arts/international/theres-more-to-winning-an-oscar-than-meets-the-eye.html

[8] Steven Yoo, Robert Kanter, David Cummings; Predicting Movie Revenue from IMDb Data, Stanford University, 2011.

[9] Jeffrey Ericson, Jesse Grodman;, A Predictor for Movie Success, Stanford University, 2013.

[10] Lyric Doshi; "Using Sentiment and Social Network Analysis to Predict Opening-Movie Box-Office Success", Massachusetts Institute of Technology, 2010

[11] Alec Kennedy; “Predicting box office success: Do critical reviews really matter?”, UC, Berkeley.